Firdavs Fayzullaev

AI/ML Engineer · LLM Systems Architect · Local Models Expert
Москва · firdavs.f.ai.dev@gmail.com · Telegram: @FirdavsAIDev

---

Summary

Архитектор production LLM-систем с глубокой экспертизой в локальных моделях и fine-tuning. Специализируюсь на RAG с персистентной памятью (Mem0/Zep), мультиагентных системах на LangGraph, оптимизации cost/latency. Создаю AI-операторов для бизнеса (15+ production Telegram-ботов) и работаю с cutting-edge технологиями voice cloning (Chatterbox). Имею мощную домашнюю лабораторию (RTX 4090, 128GB RAM) для экспериментов с локальными моделями.

Ключевые метрики

- **15+** production AI-ботов развернуто (рестораны, отели, автосалоны, B2B)
- **500K+** запросов/день обрабатывают мои системы
- **<2.0s** p95 latency на сложных agent chains
- **10+** локальных моделей настроено и оптимизировано
- **−80%** снижение затрат vs cloud API
- **MOS 4.2/5** качество voice cloning

Опыт работы

AI Engineer · Voximplant (CPaaS/UCaaS)

Январь 2024 — Настоящее время · Москва

Лидирую разработку AI-ассистента для платформы коммуникаций (15K+ бизнес-клиентов).

- **Conversational AI Platform:** Система диалоговых агентов с персистентной памятью на Mem0, обслуживающая 500K+ звонков/месяц. Метрики: p95 latency 1.8s, memory recall 94%
- **Multi-tenant RAG:** Изолированная RAG-система для B2B с гибридным поиском (pgvector + BM25 + Cohere reranker)
- **Local LLM Deployment:** Внедрил локальные модели (Llama, Qwen) для sensitive данных, настроил fine-tuning пайплайны. Снижение затрат на 80%
- **Telegram AI Bots:** Создал экосистему отраслевых ботов-операторов (10+ внедрений)

Стек: LangGraph, Mem0, pgvector, FastAPI, Kubernetes, Llama, Telegram Bot API

ML Engineer · Selector Software (AI Operations)

Март 2023 — Декабрь 2023 · Remote

Разрабатывал AI-агентов для автоматизации IT-операций и анализа логов.

- **Log Analysis Agent:** LLM-pipeline для 100GB+ логов/день. Streaming через Kafka + LangChain. Снизил MTTR на 40%
- **Runbook Automation:** Агенты с human-in-the-loop для критических операций
- **Voice Assistant PoC:** Прототип с zero-shot voice cloning для персонализированных алертов

Стек: LangChain, Zep, Kafka, ClickHouse, Grafana

AI/ML Engineer · Aimylogic (Conversational AI)

Июнь 2022 — Февраль 2023 · Москва

- Оптимизация NLU-pipeline: снизил latency на 45% через батчинг и кэширование
- Интеграция LLM (GPT) в существующую NLU-систему
- Эксперименты с локальными моделями (Llama, early Qwen) для on-premise решений

Стек: Python, Rasa, FastAPI, Redis, PostgreSQL

Ключевые проекты

🤖 Платформа AI-операторов для бизнеса (Telegram)

Разработал и развернул серию специализированных Telegram-ботов с AI для различных отраслей:

- **Рестораны:** автоматическое бронирование столиков
- **Гостиницы:** AI-консьерж для гостей
- **Автосалоны:** квалификация лидов и консультирование
- **Кафе:** прием заказов и рекомендации меню
- **B2B:** техподдержка первой линии

Все боты работают 24/7 с человекоподобными диалогами, интегрированы с CRM.

🎤 Zero-shot Voice Cloning Bot (Chatterbox)

Создал Telegram-бота с технологией zero-shot voice cloning на базе Chatterbox:

- Клонирование голоса из 5-10 секунд аудио
- Синтез речи на любом тексте клонированным голосом
- MOS 4.2/5 качество клонирования
- **Стек:** Python, Chatterbox, Telegram Bot API, FFmpeg, CUDA

🖥️ Local LLM Infrastructure

Настроил и оптимизировал локальную инфраструктуру для больших языковых моделей:

- **Модели:** Llama (8B-70B), Qwen (7B-72B), DeepSeek, Gamma (9B-27B)
- Fine-tuning и модификация манифестов под специфические задачи
- Квантизация (GGUF/GPTQ), LoRA/QLoRA адаптеры
- **Результат:** −80% cost vs API, полный контроль над данными

🚀 AI Brain для SMB процессов

Полный цикл разработки: инжест Notion/Google/Slack → гибридный RAG с цитатами → агенты на LangGraph → облако и observability.

- **Метрики:** p95 ~1.8-2.3s на 80-120k фрагментов
- **Faithfulness:** ~0.86-0.9
- **Стек:** Python, FastAPI, LangGraph, pgvector, Redis, OpenAI/Gemini

🔄 Интеграционная платформа на n8n

Десятки workflow «данные → LLM → действие»:

- Python-ноды для валидации/ETL
- Версионирование и canary-деплой при изменениях API
- **Эффект:** Снижение ручных операций на 70%

🎙️ Real-time Voice AI Platform

Полный стек voice AI: STT (Whisper) → LLM → TTS (Chatterbox):

- VAD, streaming обработка, буферизация
- Zero-shot voice cloning из 5-10 секунд
- **Метрика:** Диалоговый цикл p95 ~1.0-1.6s

Технический стек

Local Models & Fine-tuning

Llama · Qwen · DeepSeek · Gamma · Ollama · LM Studio · llama.cpp · GGUF/GPTQ · LoRA/QLoRA

Voice & Audio AI

Chatterbox · Zero-shot Cloning · Whisper · FFmpeg · PyTorch Audio · CUDA · VAD

LLM & Orchestration

LangGraph · LangChain · Semantic Kernel · n8n · CrewAI

Memory & RAG

Mem0 · Zep · pgvector · Qdrant · Weaviate · Pinecone · LlamaIndex

Bots & Integrations

Telegram Bot API · Aiogram · python-telegram-bot · Webhook · Long Polling · n8n

Cloud APIs & Services

OpenAI · Anthropic · Google Gemini · Cohere · vLLM · TGI

Infrastructure

Python · FastAPI · Docker · Kubernetes · PostgreSQL · Redis · Kafka · ClickHouse

Hardware & Optimization

NVIDIA RTX 4090 · CUDA · 128GB RAM · Quantization · Flash Attention · Model Sharding

Monitoring & DevOps

DataDog · Grafana · OpenTelemetry · Langfuse · GitHub Actions · ArgoCD

Образование

РУДН — Прикладная информатика, 2018-2022

Курсы: DeepLearning.AI (LangChain) · Cohere AI (Advanced RAG) · Anthropic (Prompt Engineering)

Языки

- **Русский:** родной
- **Английский:** техническая документация, понимание на слух
- **Узбекский:** родной

Контакты и доступность

- **Email:** firdavs.f.ai.dev@gmail.com
- **Telegram:** @FirdavsAIDev
- **LinkedIn:** /in/firdavs-fayzullaev
- **GitHub:** @firdavs-ml

- **Доступность:** Немедленно · Москва / Remote · Relocation ready

- **Оборудование:** RTX 4090, 128GB RAM, локальная инфраструктура для LLM

Ключевые достижения

- Развернул **15+ production AI-ботов** для различных бизнесов
- Обрабатываю **500K+ запросов в день** с p95 < 2s
- Настроил **10+ локальных моделей** с fine-tuning
- Снизил затраты на **80%** через локальный деплой vs cloud API
- Достиг **MOS 4.2/5** качества в voice cloning
- **−50-70%** автоматизации ручных процессов через агентов